Feature Selection Based on Fuzzy C-Means and Rough Set Theory Using Heuristic Method

Riya Jaiswal, Surendra GuptaAssoc. Professor

Department of Computer Engineering, Shri G.S. Institute of Tech. and Science23, Sir Visvesvaraya Marg, Indore (MP)

Abstract

Relevant feature selection has become an essential task to apply data mining algorithms effectively in real-world scenarios. Due to large number of features, a well known prob-lem of "noise" occurs, This problem leads to lower accuracy of machine learning classifiers due to involvement of many insignificant and irrelevant features in the dataset. Therefore, many feature selection methods have been proposed to obtain the relevant feature or feature subsets in the literature to achieve their objectives of classification and clustering. Earlier methods not explore all combinations of the features, therefore it is certain that these will fail on problems whose attributes are highly correlated resulted in poor accuracy. Therefore, a new procedure has been proposed in this work for feature selection by including the amalgamation of concepts called 'fuzzy c-means and rough set theory heuristic approach'. The work is projected in the direction of medical diagnosis to diagnose the patient's disease. In this paper medical dataset has been taken from UCI repository. A comprehensive overview, categorization, and comparison of existing feature selection methods are also done. We conclude this work with feature subsets giving better results than complete set of feature for the same algorithm. Obtained results show that the accuracy of classification is enhanced while the computational complexity has reduced.

Keywords*: *Feature Selection, Fuzzy C-Means, Rough Set Theory, Support Vector Machine

I. INTRODUCTION

The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In many real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. A detailed review of feature selection techniques devised for classification tasks can be found in (Dash and Liu, 1997).

The usefulness of a feature or feature subset is determined by both its relevancy and redundancy. A feature is said to be relevant if it is predictive of the decision feature(s), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other.

A taxonomy of feature selection approaches. Given a feature set size n, the task of FS can be seen as a search for an 'optimal' feature subset through the competing n2
candidate subsets. The definition of what an optimal subset is may vary depending on the problem to be solved. Although an exhaustive method may be used for this purpose in theory, this is quite impractical for most datasets. Usually FS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final feature subset is often reduced.(adapted from (Dash and Liu, 1997)).

Feature selection algorithms may be classified [1] into two categories based on their evaluation procedure. If an algo-rithm performs FS independently of any learning algorithm (i.e. it is a completely separatepreprocessor),

then it is a filter approach. In effect, irrelevant attributes are filtered out before induction. Filters tend to be applicable to most domains as they are not tied to any particular induction algorithm.

If the evaluation procedure is tied to the task (e.g. classi-fication [2]) of the learning algorithm, the FS algorithm em-ploys the wrapper approach. This method searches through the feature subset space using the estimated accuracy from an induction algorithm as a measure of subset suitability. Although wrappers may produce better results, they are expensive to run and can break down with very large numbers of features. This is due to the use of learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets.Therefore, in this paper, a technique has been proposed feature selection. The basic idea of this section is to reduce the features using fuzzy c-means clustering [3] with roughest

[4]theory (FCM-RS) heuristic approach. In addition, we reduce the number of features and remove the not related, unnecessary or noisy information. Besides, this improves the presentation of information prediction with speeding up the processing algorithm. To improve the prediction accuracy, use the FCM-RST [5] algorithm for feature reduction.

This paper is organized as follows. In section II an overview of work already done in this area with their advantages and disadvantages is discussed. In the section,

IIIthe proposed approach is presented. Section IV presents the details of experimental analysis of the result. Conclusion and future work is discussed in section V.

II. RELATED WORK

Substantial work has been carried out by earlier re-searchers in the field of feature selection techniques and related areas. Some of the most relevant work is discussed in this section.

Among the first rough set-based approaches is the Preset algorithm (Modrzejewski, 1993) which is another feature selector that uses rough set theory to rank heuristically the features, assuming a noise free binary domain. Since Preset does not try to explore all combinations of the features, it is certain that it will fail on problems whose attributes are highly correlated.That results in poor accuracy of classifi-cation. There have also been investigations into the use of different reduct quality measures (see (Polkowski et al, 2000) for details).

In (Zhang and Yao, 2004), a new rough set based feature selection heuristic, Parameterized Average Support Heuristic (PASH), is proposed. Unlike the existing methods, PASH is based on a special parameterized lower approximation which is defined to include all predictive instances. Predictive in-stances are instances that may produce predictive rules which hold true with a high probability but are not necessarily always true. That results in poor accuracy of classification. The traditional model could exclude predictive instances that may produce such rules. However, it requires a parameter to be defined by the user that adjusts the level of approximation.

Reducts generated from an information system are sensi-tive to changes in the system. This can be seen by removing a randomly chosen set of objects from the original object set. Those reducts frequently occurring in random subtables can be considered to be stable; it is these reducts that are encompassed by dynamic reducts [11] (Bazan et al, 1994). A disadvantage of this dynamic approach is that several subjective choices have to be made before. The dynamic reducts can be found (for instance the choice of the value of;these values are not contained in the data. Also, the huge complexity of finding all reducts within subtables forces the use of other methods.

In (Han et al., 2004), a feature selection method based on an alternative dependency measure is presented. The technique was originally proposed to avoid the calculation of discernibility functions or positive regions, which can be computationally expensive without optimizations.
The approaches reported in (Bjorvand and Komorowski, 1997; Wroblewski,´ 1995) use genetic algorithms to discover optimal or close-to-optimal reducts. Reduct candidates are encoded as bit strings, with the value in

position i set if the ith attribute is present. The fitness function depends on two parameters. The first is the number of bits set. The function penalises those strings which have larger numbers of bits set, driving the process to find smaller reducts. The reduct should discern between as many objects as possible (ideally all of them). Although this approach to FS is not guaranteed to find minimal subsets, it may find many subsets for any given dataset. It is also useful for situations where new objects are added to or old objects are removed from a dataset the reducts [12] generated previously can be used as the initial population for the new reduct-determining process. The main drawback is the time taken to compute each bit string's fitness, which is O(a*o2), where a is the number of attributes and o the number of objects in the dataset. The extent to which this hampers performance depends mainly on the population size.

Effective hybrid attribute reduction algorithm based on a generalized fuzzy rough model. A theoretic framework of fuzzy rough model based on fuzzy relations is accessi-ble, which contain a foundation for algorithm construction. Though, several attributes derive significance measures based on the proposed fuzzy-rough model and create a forward greedy algorithm for hybrid attribute reduction.

(Multivariate) Error-Weighted Uncorrelated Shrunken Centroid (EWUSC): this method is based on the uncorrelated shrunken centroid (USC) and shrunken centroid (SC). The shrunken centroid is found by dividing the average gene expression for each gene in each class by the standard deviation for that gene in the same class. This way higher weight is given to genes whose expression is the same among different samples in the same class. New samples are assigned to the label with the nearest average pattern (using squared distance). The uncorrelated shrunken centroid approach removes redundant features by finding genes that are highly correlated in the set of genes already found by SC. The EWUSC uses both of these steps and in addition adds error-weights (based on within-class variability) so that noisy genes will be downgraded and redundant genes are removed. The algorithms perform well when the number of relevant genes is less than 1000.

Statistical methods often assume a Gaussian distribution on the data. The central limit theorem can guarantee that large datasets are always normally distributed. Even though all these methods can be highly accurate in classifying information there is no biological significance proven with the genes that are identified by them. None of the above methods have indicated whether the results are actually biologically relevant or not. In addition filter methods are generally faster than wrappers but do not take into account the classifier which can be a disadvantage.

Leave-one-out calculation sequential forward selection (LOOCSFS) [13] is a very widely used feature selection method for cancer data based on sequential forward selection (SFS). It adds features in an initially empty set and calculates the leave-one-out cross-validation error. It can be used in combination with a recursive support vector machine (R-SVM) algorithm that selects important feature. The contri-bution factor, based on minimal error of the support vector machine, of each feature is calculated and ranked. The top ranked feature are chosen for the subset. The number of the feature in the feature subset for both LOOCSFS and GLGS has to be given in advance which can be a disadvantage since the most important feature are not known in advance.

But the existing methods are not focusing on reducing ir-relevant features which improves the classification accuracy.

III.PROPOSED METHOD

The proposed approach is focused on reducing irrelevant features so that algorithms can work on important features only there is no need to feed all the features to the required algorithm.

It consists of data preprocessing module, feature extraction module, feature selection module, classification module, per-formance result module. Architecture of the proposed system is illustrated in figure I.

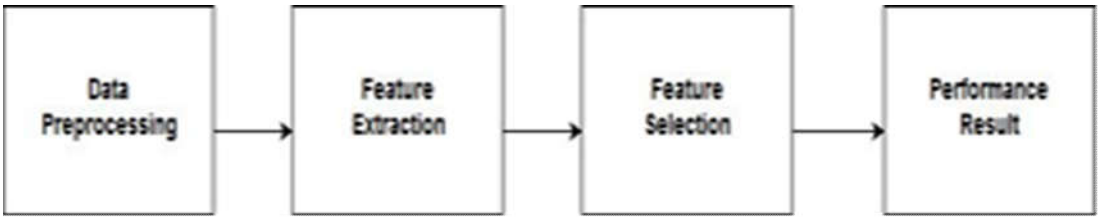

Fig. 1: Architecture Diagram of Proposed Approach

A. Data preprocessing Module

The system takes medical dataset as an input. It read the image of dataset, Resize the dataset and Filter the dataset.

- Resized image = imresize (Input image,[n,m]). Because Larger images take up more memory, so if you're never going to use them beyond a certain resolution, it doesn't make any sense to waste memory space.
- Filter image = median filter. Because the median filter is used to reduce noise in an image, However, it often does a better job than the mean filter of preserving useful detail in the image.

B. Feature Extraction Module

- Apply Feature Extraction on the given data to extract features and then apply feature selection.
- Apply Histogram of Oriented Gradient[g_x,g_y,],g_x is displacement corresponding x-axis,g_y is displacement corresponding y-axis,is angle for every pixel that is g_y/g_x.

HOG is used because HOG-based classifiers are ex-tremely fast to train and evaluate, which enables confi-dence level estimation.HOG is most often used as a feature (or part of a feature) for image recognition problems. HOG is much easier and quicker to compute. Furthermore, HOG is used to describe a whole image / image patch.

- Extract features.

C. Feature Selection Module

Feature selection method seek to reduce the number of attributes in the dataset. By applying rough set theory the proposed approach reduce feature but it retains the originality of data.

- Fuzzy C-Means

Use FCM (Fuzzy c-Means) algorithm.
Apply Fuzzy c-Means Clustering [8] method on the com-puted features.
The output should be a centroid.

- Rough Set Theory (Heuristic Approach)

Now apply rough set theory over all the computed centres. RST [6] will give the best minimal subset which retains the accuracy of original set.
Pass the output of FCM (Fuzzy c-Means) [9] i.e., centroid as input parameter.

Reducts: Is the minimum no of attributes that preserves the its indiscernibility relationship.
Core: It form indiscernibility relation. Those attribute which has same values but different decision attribute values.
Algorithm 1 Heuristic Algorithm

Input: Extracted Feature.

Output: Selected Feature.
Procedure:
1: Loop
2: Remove attribute 'A' and traverse the dataset(each tuple) for checking whether 'A' is part of 'core' or not.
3: if Core then
4: Add To (core attribute list)
5: if Reject and continue to pick next attribute.Remove all attributes that are not in the 'core' then
6: Now, combine (core attribute list) with decision attribute to find reduct.
7: if noreduct making then
8: Add tuples 1 by 1 to form reduct.(to make rules it should consistant)
9: ComapareReducts : Choose those reducts which have higher support and confidence.

D. Classification Module

1) Load the selected feature.
2) Use SVM (Support Vector Machine) for classification. Works well for high dimensional data. The SVM can betrained to classify both linearly separable and non-linearly separable data. Represents decision boundary using subset of training examples (from both classes), known as support vectors.
Characteristics of SVM : The SVM [7] learning problem can be formulated as a convex optimization problem, in which efficient algorithm are available to find global min-imum of objective function.SVM performs capacity control by maximizing the margin of decision boundary.SVM [10] can be applied to categorical data by introducing dummy variable for each categorical attribute value present in the data.

Properties of SVM : Flexibility in choosing a similarity function. Sparseness of solution when dealing with large data sets. Ability to handle large feature space. Nice math property: a simple convex optimization problem which is guaranteed to converge to a single global solution.

Classification module takes resultant dataset as an input. Dataset is divided into two parts: training set and testing set for applying classification. Classifier is trained using training set and using testing set its accuracy is computed.

E. Performance Evaluation Module

There are many parameters available which can be used to check performance of classification. The parameters used in the project to check classification performance are :
Accuracy : It is the ratio of number of correct prediction to total no of prediction.
Sensitivity : It is considered as a measure of complete-ness. It tells about how many positive class samples are correctly classified.
Specificity : Specificity is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).
Precision : It is a measure of exactness. It can be defined as, from the samples labelled as positive, how many actually belongs to positive class.
Recall : Recall = sensitivity
F-measure : It is the harmonic mean between recall and precision.
Overlap : To cover something partly by going over its edge; to cover part of the same space.

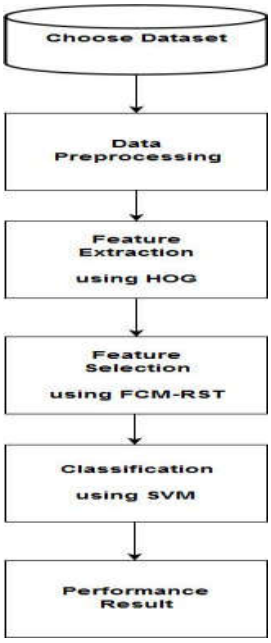

Fig. 2: Flow Diagram of Proposed Approach

IV. TESTING AND RESULTS

Proposed approach is implemented using Matlab. It is evaluated on Lung Breast and Brain dataset which is taken from UCI repository.

A comparison is made on proposed model by applying classifiers namely SVM between original feature and selected feature.

TABLE I: CLUSTER CENTER VALUE FOR EACH CLUSTER OF THE HIGH DIMENSIONAL DATASETS

Dataset	Centroids
Lung	0.6831
Breast	0.8035
Brain	0.6081

TABLE II: OPTIMIZED FEATURE VALUES FROM FEATURE SELECTION TECHNIQUE ROUGH SET THEORY USING HEURISTIC APPROACH

Dataset	Centroids	OF's Original Features Total	SF's Selected Features Total
Lung	0.6831	4000	76
Breast	0.8035	4000	75
Brain	0.6081	4000	76

TABLE III: THE COMPARISION OF CLASSIFICA-TION ACCURACIES WITH DIFFERENT DATASETS

Parameters	Lung Result	Breast Result	Brain Result
Accuracy	.93	.91	.94
Sensitivity	.92	.90	.93
Specificity	.89	.86	.89
Precision	.91	.89	.93
Recall	.92	.90	.93
F-measure	.93	.91	.91
Overlap	.87	.85	.85

V. ANALYSIS OF THE RESULT

From the tables III, it is observed that classification accuracy is better for the proposed model.

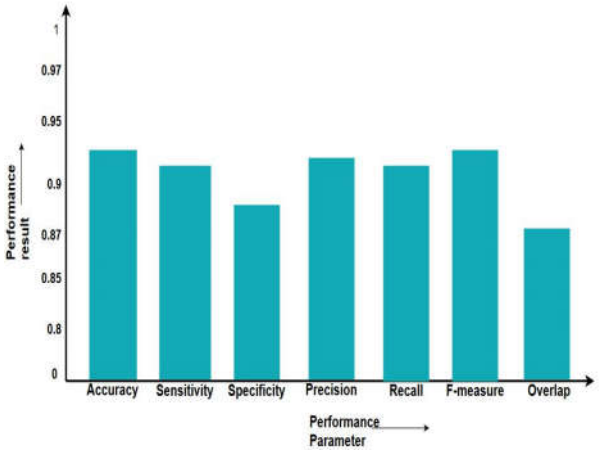

Fig. 3: Performance evaluation using various parameter measures [Lung Dataset]

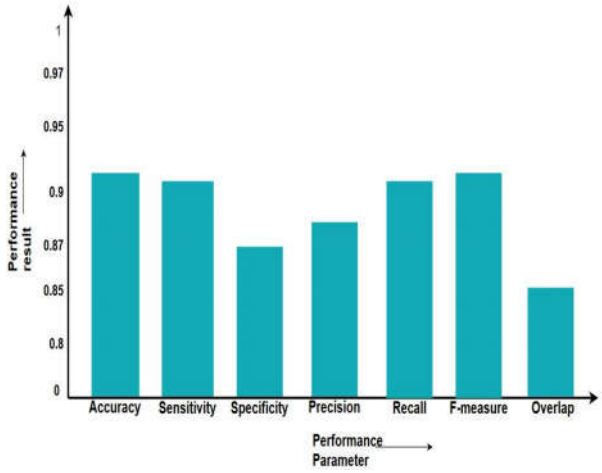

Fig. 4: Performance evaluation using various parameter measure [Breast Dataset]

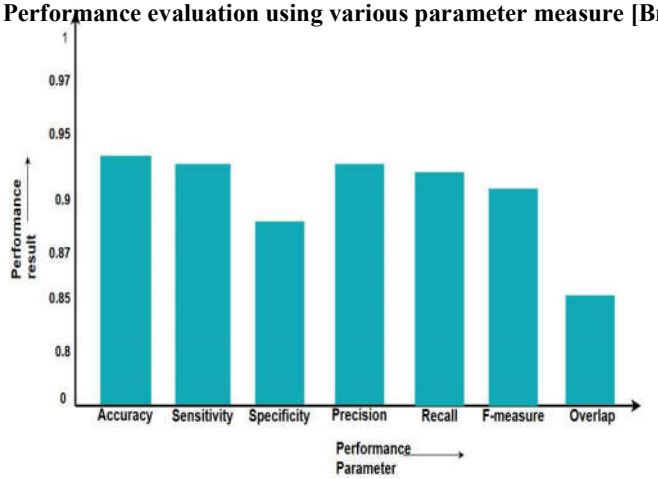

Fig. 5: Performance evaluation using various parameter measures [Brain Dataset]

VI. CONCLUSION

Accuracy is most significant in the field of medical diagnosis to diagnose the patient's disease. Experimental results show that Feature Selection, a Preprocessing tech-nique greatly boost the accuracy of classification.For data mining and machine learning problems Feature selection, as an important data preprocessing strategy, has been proven to be effective and efficient in preparing high-dimensional data. The objectives of feature selection [14] include building simpler and more comprehensible models, improving data mining performance, and preparing clean, understandable data. Once the best feature selection method [15] is identified for a particular dataset the same can be used to enhance the classifier accuracy.The proposed system which shows the better performance than the existing feature selection process.

REFERENCES

[1] K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988
[2] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification and Scene Analysis. New York: Wiley, 1999.
[3] F. Masulli and S. Rovetta, "Soft transition from probabilistic to possi-bilistic fuzzy clustering," IEEE Trans. Fuzzy Syst., vol. 14, no. 4, pp. 516–527, Aug. 2006.
[4] P. Lingras and C. West, "Interval set clustering of web users with rough K-means," J. Intell. Inf. Syst., vol. 23, no. 1, pp. 5–16, 2004.
[5] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," Int. J. Gen. Syst., vol. 17, no. 2/3, pp. 191– 209, 1990.
[6] A.Pethalakshmi ,A.Banumathi "A novel approach in clustering via rough set.Internatonal journal of Science and Research, Volume 2 Issue 7, pp.139-145,July 2013
[7] Y.L.Zhang, N.Guo, H.Du and W.H Li, "Automated defect recognition of C- SAM images in IC packaging using Support Vector Machines," The International Journal of Advanced Manufacturing Technology 25, 1191 - 1196,2005.
[8] R. Krishnapuram and J. M. Keller, "A possibilistic approach to cluster-ing," IEEE Trans. Fuzzy Syst., vol. 1, no. 2, pp. 98–110, May 1993.
[9] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," IEEE Trans. Fuzzy Syst., vol. 13, no. 4, pp. 517–530, Aug. 2005.
[10] A.MehmetFatih, "Support vector machines combined with feature se-lection for breast cancer diagnosis," Expert Systems with Applications, Vol. 36, pp.32403247,2009.

[11] Bazan, J., Skowron, A., and Synak, P. (1994). Dynamic reducts as a tool for extracting laws from decision tables. In Proceedings of the 8th Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence 869, Springer-Verlag, pp. 346–355.
[12] Starzyk, J. A., Nelson, D. E., and Sturtz, K. (2000). A Mathematical Foundation for Improved Reduct Generation in Information Systems. Journal of Knowledge and Information Systems, Vol. 2, No. 2, pp.131-146.
[13] Hoos, H. H., and Stutzle,¨ T. (1999). Towards a Characterisation of the Behaviour of Stochastic Local Search Algorithms for SAT. Artificial Intelligence, Vol. 112, pp. 213–232.
[14] Jensen, R., Shen, Q., and Tuson, A. (2005). Finding Rough Set Reducts with SAT. In Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, LNAI 3641, pp. 194-203.
[15] Nguyen, S. H., and Nguyen, H. S. (1996). Some efficient algorithms for rough set methods. In Proceedings of the Conference of Informa-tion Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 1451–1456.